Homework 2: Microcensus

NAME

2025-11-16

Contents

Load Data and Setup to identify key variables

```
df <- haven::read_dta("10752_da10_de_v4_0.dta")

person_id_var <- "asbper"
household_id_var <- "asbhh"
quarter_var <- "aquartal"
year_var <- "ajahr"

df <- df %>% mutate(.quarter=.data[[quarter_var]])
```

1. Household vs. Person

(a) For each quarter and for the whole year, show the unweighted and weighted number of households and persons using both survey weight variables (*gew1* and *gewjahr*), and explain the difference between these weights and when to use each of them.

```
# Prepare household-quarter level dataset
# (one row per household-quarter observation)
if (!is.na(household_id_var)) {
  hh_q <- df %>%
    filter(!is.na(.quarter)) %>%
    group_by(.quarter, hh = .data[[household_id_var]]) %>%
    slice(1) %>%
    ungroup()
} else {
  # fallback: treat each row as household occurrence if household id unknown
  hh_q <- df %>% mutate(hh = row_number()) %>% group_by(.quarter, hh) %>%
    slice(1) %>% ungroup()
}

# Unweighted counts by quarter
unweighted_q <- df %>%
  group_by(.quarter) %>%
  summarise(
    n_persons = n(),
    n_households = n_distinct(.data[[household_id_var]] %||% row_number())
  )

# Weighted persons by quarter using gew1 (person-level weight)
weighted_persons_q <- df %>%
  group_by(.quarter) %>%
  summarise(
    persons_w_gew1 = sum(gew1, na.rm = TRUE),
    persons_w_gewjahr = sum(gewjahr, na.rm = TRUE)
  )

# Weighted households by quarter
# take first observation of each household-quarter and sum its weight
weighted_households_q <- hh_q %>%
  summarise(
    households_w_gew1 = sum(gew1, na.rm = TRUE),
    households_w_gewjahr = sum(gewjahr, na.rm = TRUE)
  ) %>%
  bind_cols(. )

unweighted_q; weighted_persons_q; weighted_households_q
```

```
## # A tibble: 4 x 3
##   .quarter n_persons n_households
##      <dbl>     <int>        <int>
## 1        1     43281        20234
## 2        2     43444        20263
## 3        3     43651        20311
## 4        4     43739        20390

## # A tibble: 4 x 3
##   .quarter persons_w_gew1 persons_w_gewjahr
##      <dbl>          <dbl>             <dbl>
## 1        1       8789980.          2197546.
## 2        2       8795767.          2198997.
## 3        3       8804959.          2201293.
## 4        4       8837697.          2209477.

## # A tibble: 1 x 2
##   households_w_gew1 households_w_gewjahr
##               <dbl>                <dbl>
## 1         16078571.             4019744.
```

```r
# Whole year totals (use gewjahr for year-level weights)
total_unweighted <- tibble(
  n_persons = nrow(df),
  n_households = n_distinct(df[[household_id_var]] %||% seq_len(nrow(df)))
)

total_weighted <- tibble(
  persons_w_gew1 = sum(df$gew1, na.rm = TRUE),
  persons_w_gewjahr = sum(df$gewjahr, na.rm = TRUE),
  households_w_gew1 = df %>% group_by(.data[[household_id_var]]) %>%
    slice(1) %>% summarise(w = sum(gew1, na.rm = TRUE)) %>% pull(w) %>%
    sum(na.rm = TRUE),
  households_w_gewjahr = df %>% group_by(.data[[household_id_var]]) %>%
    slice(1) %>% summarise(w = sum(gewjahr, na.rm = TRUE)) %>% pull(w) %>%
    sum(na.rm = TRUE)
)

total_unweighted; total_weighted
```

```
## # A tibble: 1 x 2
##   n_persons n_households
##       <int>        <int>
## 1    174115        33555
```

```
## # A tibble: 1 x 4
##   persons_w_gew1 persons_w_gewjahr households_w_gew1 households_w_gewjahr
##            <dbl>             <dbl>             <dbl>                <dbl>
## 1      35228403.          8807314.          6651560.             1662931.
```

Unweighted counts: 174115 persons and 33555 households observed in the sample.

Weighted (annual) counts using *gewjahr*: 8,807,314 persons and 1,662,931 households (population projection).

gew1 is the survey weight appropriate for quarter-level analyses; *gewjahr* is the annual weight that adjusts observations to represent the whole year. Use *gew1* when analyzing quarter-specific quantities; use *gewjahr* for yearly aggregates or when summarizing across quarters.

(b) Compare the gender distribution (bsex) of household heads (household reference person: xhrp == 1) to the overall gender distribution in the sample, and explain why this comparison is relevant for interpreting household-level statistics.

```r
heads <- df %>% filter(xhrp == 1)

# Unweighted distributions
dist_heads_unw <- heads %>% count(bsex) %>% mutate(pct = n / sum(n))
dist_overall_unw <- df %>% count(bsex) %>% mutate(pct = n / sum(n))

# Weighted distributions (person-level weights)
dist_heads_w <- heads %>% group_by(bsex) %>%
                summarise(w = sum(gew1, na.rm = TRUE)) %>%
                mutate(pct = w / sum(w))
```

```r
dist_overall_w <- df %>% group_by(bsex) %>%
  summarise(w = sum(gew1, na.rm = TRUE)) %>% mutate(pct = w / sum(w))

dist_heads_unw; dist_overall_unw; dist_heads_w; dist_overall_w
```

```
## # A tibble: 2 x 3
##   bsex              n   pct
##   <dbl+lbl>     <int> <dbl>
## 1 1 [Männlich] 48757 0.600
## 2 2 [Weiblich] 32441 0.400

## # A tibble: 2 x 3
##   bsex              n   pct
##   <dbl+lbl>     <int> <dbl>
## 1 1 [Männlich] 84424 0.485
## 2 2 [Weiblich] 89691 0.515

## # A tibble: 2 x 3
##   bsex                 w   pct
##   <dbl+lbl>        <dbl> <dbl>
## 1 1 [Männlich] 9774212. 0.608
## 2 2 [Weiblich] 6304358. 0.392

## # A tibble: 2 x 3
##   bsex                  w   pct
##   <dbl+lbl>         <dbl> <dbl>
## 1 1 [Männlich] 17370261. 0.493
## 2 2 [Weiblich] 17858142. 0.507
```

Household heads (weighted): 60.8% male; Overall (weighted): 49.3% male.

Household heads are more likely to be male than the population average; household-level indicators may therefore reflect male-dominated characteristics (e.g., employment rates, hours).

(c) Calculate the household-level average and the individual-level average of hours worked (*dstd*) and compare these averages by gender; interpret the differences you observe.

```r
# Individual-level average hours by gender
indiv_avg <- df %>% group_by(bsex) %>%
  summarise(indiv_avg_dstd = mean(dstd, na.rm = TRUE))


# Household-level average: first compute household-level average hours
# create an explicit `hh` column so joins are straightforward
hh_avg <- df %>%
  group_by(hh = .data[[household_id_var]]) %>%
  summarise(hh_avg_dstd = mean(dstd, na.rm = TRUE)) %>%
  ungroup()

# join head gender to have household-level averages by household head gender
hh_with_head <- df %>%
  filter(xhrp == 1) %>%
  select(hh = .data[[household_id_var]], head_bsex = bsex) %>%
  distinct()

# join by `hh`
hh_gender_avg <- hh_avg %>% left_join(hh_with_head, by = "hh")

indiv_avg; head_gender_avg <- hh_gender_avg %>% group_by(head_bsex) %>%
  summarise(mean_hh_avg = mean(hh_avg_dstd, na.rm = TRUE)); head_gender_avg
```

```
## # A tibble: 2 x 2
##   bsex          indiv_avg_dstd
##   <dbl+lbl>              <dbl>
## 1 1 [Männlich]            19.7
## 2 2 [Weiblich]            12.5
```

```
## # A tibble: 2 x 2
##   head_bsex    mean_hh_avg
##   <dbl+lbl>          <dbl>
## 1 1 [Männlich]        17.7
## 2 2 [Weiblich]        13.6
```

Men work more hours on average at the individual level (19.7 vs 12.5). Households headed by men also show higher household-average hours (17.7 vs 13.6), but the smaller household-level gap suggests composition effects (e.g., number of workers per household and selection into headship).

2. Panel structure

(a) Compute the weighted share of individuals who are observed at least twice within the year, and compute the weighted share of households that have at least one person observed at least twice.

```
# Count observations per person within the year (create explicit `pid`)
person_counts <- df %>%
  group_by(pid = .data[[person_id_var]]) %>%
  summarise(n_obs = n(), w = sum(gewjahr, na.rm = TRUE), .groups = "drop")

persons_at_least_two <- person_counts %>% filter(n_obs >= 2)

# Weighted share (using annual weight)
share_persons_ge2 <- sum(persons_at_least_two$w, na.rm = TRUE) /
  sum(person_counts$w, na.rm = TRUE)

# Households with at least one person observed twice
persons_flag <- person_counts %>% mutate(flag_ge2 = n_obs >= 2)

# Create household-person links with explicit names `hh` and `pid`
hh_persons <- df %>% select(hh = .data[[household_id_var]],
                            pid = .data[[person_id_var]]) %>% distinct()
```

```r
# Join flags and aggregate to household level
hh_flag <- hh_persons %>%
  left_join(persons_flag, by = "pid") %>%
  group_by(hh) %>%
  summarise(any_person_ge2 = any(flag_ge2, na.rm = TRUE), .groups = "drop")

# compute weighted household share
#use household annual weight from first observation
hh_w <- df %>% group_by(hh = .data[[household_id_var]]) %>% slice(1) %>%
  ungroup() %>% select(hh, gewjahr)

hh_with_w <- hh_flag %>% left_join(hh_w, by = "hh")
share_households_with_person_ge2 <-
  sum(hh_with_w$gewjahr[hh_with_w$any_person_ge2], na.rm = TRUE) /
  sum(hh_with_w$gewjahr, na.rm = TRUE)

share_persons_ge2; share_households_with_person_ge2
```

```
## [1] 0.8785494
```

```
## [1] 0.722392
```

Weighted share of individuals observed at least twice: 87.85%. Weighted share of households with at least one person observed at least twice: 72.24%.

A higher share of repeat observations increases the potential for panel analyses (tracking individuals or households over time). If the share is low, it is required to interpret longitudinal findings cautiously due to small effective sample sizes.

(b) Restrict the dataset to individuals observed in Q1 and Q2 and aged between 19 and 64 (*xbalt5*): (i) how many persons switch employment status (*xlfi*) between Q1 and Q2 and in which directions? Report the weighted shares of these switching groups among all persons in Q1. (ii) Focus on one switching group and describe at least two additional characteristics (variables not used in class) comparing this group to the Q1 sample or another appropriate comparison group; explain what you find.

```r
# xlfi: employment status
# 1 employed, 2 unemployed, 3 non-employed persons (incl. military/civil service)

# Standardise quarter labelling: try to map common encodings to 1 and 2
df <- df %>% mutate(.qnum = case_when(
  .quarter == 1 ~ 1,
  .quarter == 2 ~ 2,
  TRUE ~ NA_real_
))

q1q2 <- df %>% filter(.qnum %in% c(1,2), xbalt5 == 1)

# keep only persons observed in both Q1 and Q2
persons_q1q2 <- q1q2 %>% group_by(pid = .data[[person_id_var]]) %>%
  filter(n_distinct(.qnum) == 2)
```

```
# create wide table for xlfi (employment status) in Q1 and Q2
xlfi_wide <- persons_q1q2 %>%
  select(pid = .data[[person_id_var]], .qnum, xlfi, gewjahr) %>%
  pivot_wider(names_from = .qnum, values_from = c(xlfi, gewjahr),
              names_sep = "_")

# number of persons who switch between categories
switches <- xlfi_wide %>% filter(!is.na(xlfi_1) &
                                   !is.na(xlfi_2) & xlfi_1 != xlfi_2)
switch_counts <- switches %>% count(xlfi_1, xlfi_2)

sw <- xlfi_wide %>% filter(!is.na(xlfi_1) & !is.na(xlfi_2) &
                             xlfi_1 != xlfi_2) %>%
      group_by(xlfi_1, xlfi_2) %>%
      summarise(w = sum(gewjahr_1, na.rm = TRUE), .groups = "drop")
    total_q1_w <- sum(xlfi_wide$gewjahr_1, na.rm = TRUE)
    sw <- sw %>% mutate(pct = 100 * w / total_q1_w)

switch_counts; sw
```

```
## # A tibble: 6 x 3
##   xlfi_1                                    xlfi_2                             n
##   <dbl+lbl>                                 <dbl+lbl>                      <int>
## 1 1 [Erwerbstätige]                         2 [Arbeitslose]                   11
## 2 1 [Erwerbstätige]                         3 [Nicht-Erwerbspersonen (ink~    28
## 3 2 [Arbeitslose]                           1 [Erwerbstätige]                 18
## 4 2 [Arbeitslose]                           3 [Nicht-Erwerbspersonen (ink~    15
## 5 3 [Nicht-Erwerbspersonen (inkl Präs/Ziv)] 1 [Erwerbstätige]                 70
## 6 3 [Nicht-Erwerbspersonen (inkl Präs/Ziv)] 2 [Arbeitslose]                   38
```

```
## # A tibble: 6 x 4
##   xlfi_1                                    xlfi_2                       w   pct
##   <dbl+lbl>                                 <dbl+lbl>                <dbl> <dbl>
## 1 1 [Erwerbstätige]                         2 [Arbeitslose]            561. 0.722
## 2 1 [Erwerbstätige]                         3 [Nicht-Erwerbspersone~ 1256. 1.62
## 3 2 [Arbeitslose]                           1 [Erwerbstätige]        1005. 1.29
## 4 2 [Arbeitslose]                           3 [Nicht-Erwerbspersone~  681. 0.876
## 5 3 [Nicht-Erwerbspersonen (inkl Präs/Ziv)] 1 [Erwerbstätige]        3553. 4.57
## 6 3 [Nicht-Erwerbspersonen (inkl Präs/Ziv)] 2 [Arbeitslose]          2155. 2.77
```

We find the largest flows from non-employed to employed (4.57%).

```
# Analyze switching group xlfi_1 == 3 -> xlfi_2 == 1
# compare xmigr / xmigr_gen to Q1 sample
q1_attrs <- df %>%
  filter(.qnum == 1) %>%
  transmute(pid = .data[[person_id_var]],
            xmigr = xmigr, xmigr_gen = xmigr_gen) %>%
  distinct()

# attach Q1 attributes to wide table (weights from Q1: gewjahr_1)
xlfi_wide2 <- xlfi_wide %>% left_join(q1_attrs, by = "pid")
```

```r
# define switchers and totals
switchers <- xlfi_wide2 %>% filter(xlfi_1 == 3 & xlfi_2 == 1)
total_q1_w <- sum(xlfi_wide2$gewjahr_1, na.rm = TRUE)
switch_w <- sum(switchers$gewjahr_1, na.rm = TRUE)
pct_switch_of_q1 <- 100 * switch_w / total_q1_w

# Weighted distribution of xmigr: overall Q1 vs switchers
overall_xmigr <- xlfi_wide2 %>%
  filter(!is.na(xmigr)) %>%
  group_by(xmigr) %>%
  summarise(w = sum(gewjahr_1, na.rm = TRUE), .groups = "drop") %>%
  mutate(pct = 100 * w / sum(w))

switch_xmigr <- switchers %>%
  filter(!is.na(xmigr)) %>%
  group_by(xmigr) %>%
  summarise(w = sum(gewjahr_1, na.rm = TRUE), .groups = "drop") %>%
  mutate(pct = 100 * w / sum(w))

# Weighted distribution of xmigr_gen: overall Q1 vs switchers
overall_xmigr_gen <- xlfi_wide2 %>%
  filter(!is.na(xmigr_gen)) %>%
  group_by(xmigr_gen) %>%
  summarise(w = sum(gewjahr_1, na.rm = TRUE), .groups = "drop") %>%
  mutate(pct = 100 * w / sum(w))

switch_xmigr_gen <- switchers %>%
  filter(!is.na(xmigr_gen)) %>%
  group_by(xmigr_gen) %>%
  summarise(w = sum(gewjahr_1, na.rm = TRUE), .groups = "drop") %>%
  mutate(pct = 100 * w / sum(w))

# Print concise tables and a short interpretation
cat("Summary: switching group xlfi 3 -> 1\n")
```

```
## Summary: switching group xlfi 3 -> 1
```

```r
cat(sprintf("Switchers represent %.2f%% of the Q1 sample (weighted).\n\n",
            pct_switch_of_q1))
```

```
## Switchers represent 4.57% of the Q1 sample (weighted).
```

```r
cat("Overall Q1 xmigr (0=no,1=yes):\n"); print(overall_xmigr)
```

```
## Overall Q1 xmigr (0=no,1=yes):
```

```
## # A tibble: 2 x 3
##   xmigr                                w   pct
##   <dbl+lbl>                        <dbl> <dbl>
## 1 0 [Kein Migrationshintergrund] 57037.  73.4
## 2 1 [Migrationshintergrund]      20642.  26.6
```

```r
cat("\nSwitchers xmigr (0=no,1=yes):\n"); print(switch_xmigr)
```

```
##
## Switchers xmigr (0=no,1=yes):

## # A tibble: 2 x 3
##   xmigr                                w   pct
##   <dbl+lbl>                        <dbl> <dbl>
## 1 0 [Kein Migrationshintergrund] 2749.  77.4
## 2 1 [Migrationshintergrund]       805.  22.6
```

```r
cat("\nOverall Q1 xmigr_gen (0=no,1=first gen,2=second gen):\n");
```

```
##
## Overall Q1 xmigr_gen (0=no,1=first gen,2=second gen):
```

```r
print(overall_xmigr_gen)
```

```
## # A tibble: 3 x 3
##   xmigr_gen                             w   pct
##   <dbl+lbl>                         <dbl> <dbl>
## 1 0 [Ohne Migrationshintergrund] 57037. 73.4
## 2 1 [Erste Generation]            6711.  8.64
## 3 2 [Zweite Generation]          13931. 17.9
```

```r
cat("\nSwitchers xmigr_gen (0=no,1=first gen,2=second gen):\n");
```

```
##
## Switchers xmigr_gen (0=no,1=first gen,2=second gen):
```

```r
print(switch_xmigr_gen)
```

```
## # A tibble: 3 x 3
##   xmigr_gen                            w   pct
##   <dbl+lbl>                        <dbl> <dbl>
## 1 0 [Ohne Migrationshintergrund] 2749. 77.4
## 2 1 [Erste Generation]            144.  4.06
## 3 2 [Zweite Generation]           661. 18.6
```

Migration background (xmigr=1): Q1 = 26.6%; switchers = 22.6%. By generation (xmigr_gen): Q1 first-gen = 8.6%, second-gen = 17.9%; switchers first-gen = 4.1%, second-gen = 18.6%. Switchers are relatively less likely to have a migration background than the Q1 sample overall. Job entry is concentrated among second-generation migrants in comparison to first-generation migrants.

3. Housing

(a) Restrict the data to the household level and report the weighted share of households that own their home, rent, or live for free (*wrechtk*), separately by urbanisation degree (*xurb*) and by Bundesland (*xnuts2*).

```
# Restrict to household-level
hh <- df %>% group_by(hh = .data[[household_id_var]]) %>% slice(1) %>% ungroup()

table_housing <- hh %>%
  group_by(xurb, xnuts2, wrechtk) %>%
  summarise(weight = sum(gewjahr, na.rm = TRUE)) %>%
  group_by(xurb, xnuts2) %>%
  mutate(pct = weight / sum(weight)) %>%
  ungroup()

table_housing
```

```
## # A tibble: 66 x 5
##    xurb                          xnuts2              wrechtk         weight    pct
##    <dbl+lbl>                     <dbl+lbl>           <dbl+lbl>        <dbl>  <dbl>
##  1 1 [hohe Bevölkerungsdichte] 13 [Wien]           1 [zur Miete]   3.00e5 0.776
##  2 1 [hohe Bevölkerungsdichte] 13 [Wien]           2 [im Eigentum] 7.58e4 0.196
##  3 1 [hohe Bevölkerungsdichte] 13 [Wien]           3 [mietfrei od~ 1.07e4 0.0278
##  4 1 [hohe Bevölkerungsdichte] 21 [Kärnten]        1 [zur Miete]   1.16e4 0.557
##  5 1 [hohe Bevölkerungsdichte] 21 [Kärnten]        2 [im Eigentum] 8.22e3 0.395
##  6 1 [hohe Bevölkerungsdichte] 21 [Kärnten]        3 [mietfrei od~ 9.95e2 0.0479
##  7 1 [hohe Bevölkerungsdichte] 22 [Steiermark]     1 [zur Miete]   3.31e4 0.577
##  8 1 [hohe Bevölkerungsdichte] 22 [Steiermark]     2 [im Eigentum] 2.17e4 0.379
##  9 1 [hohe Bevölkerungsdichte] 22 [Steiermark]     3 [mietfrei od~ 2.55e3 0.0445
## 10 1 [hohe Bevölkerungsdichte] 31 [Oberösterreich] 1 [zur Miete]   3.25e4 0.777
## # i 56 more rows
```

Table below is a simplified version grouped by home ownership status only.

```
table_housing %>%
  group_by(wrechtk) %>% summarise(w = sum(weight, na.rm = TRUE),
                                  .groups = "drop") %>%
  mutate(pct = 100 * w / sum(w))
```

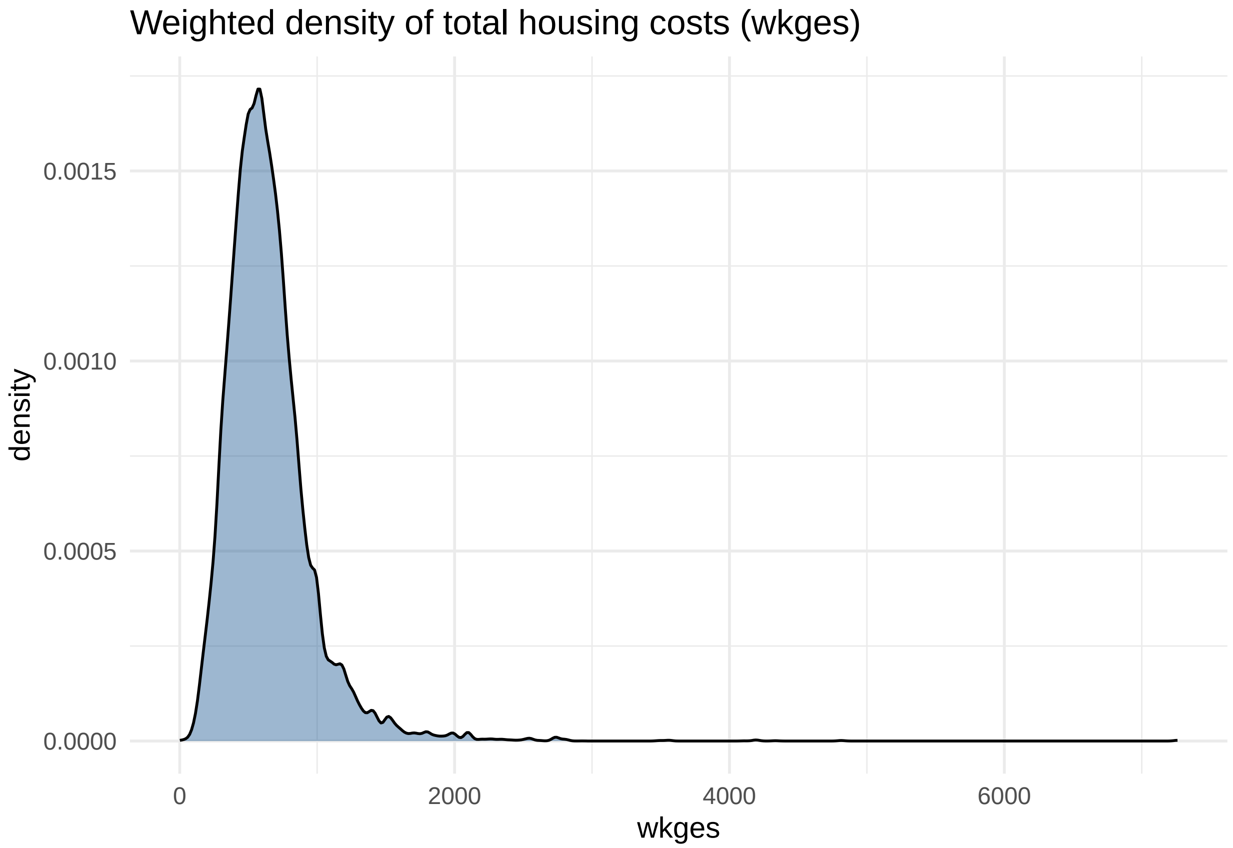
```
## # A tibble: 3 x 3
##   wrechtk                                w   pct
##   <dbl+lbl>                          <dbl> <dbl>
## 1 1 [zur Miete]                    726849. 43.7
## 2 2 [im Eigentum]                  805762. 48.5
## 3 3 [mietfrei oder unentgeltlich] 130320.  7.84
```

(b) Restrict to households renting as main tenants and construct two housing cost variables — total housing costs (*wkges*) and cost per square meter (*wm2*) — then plot the weighted distributions of these two variables and interpret what you observe.

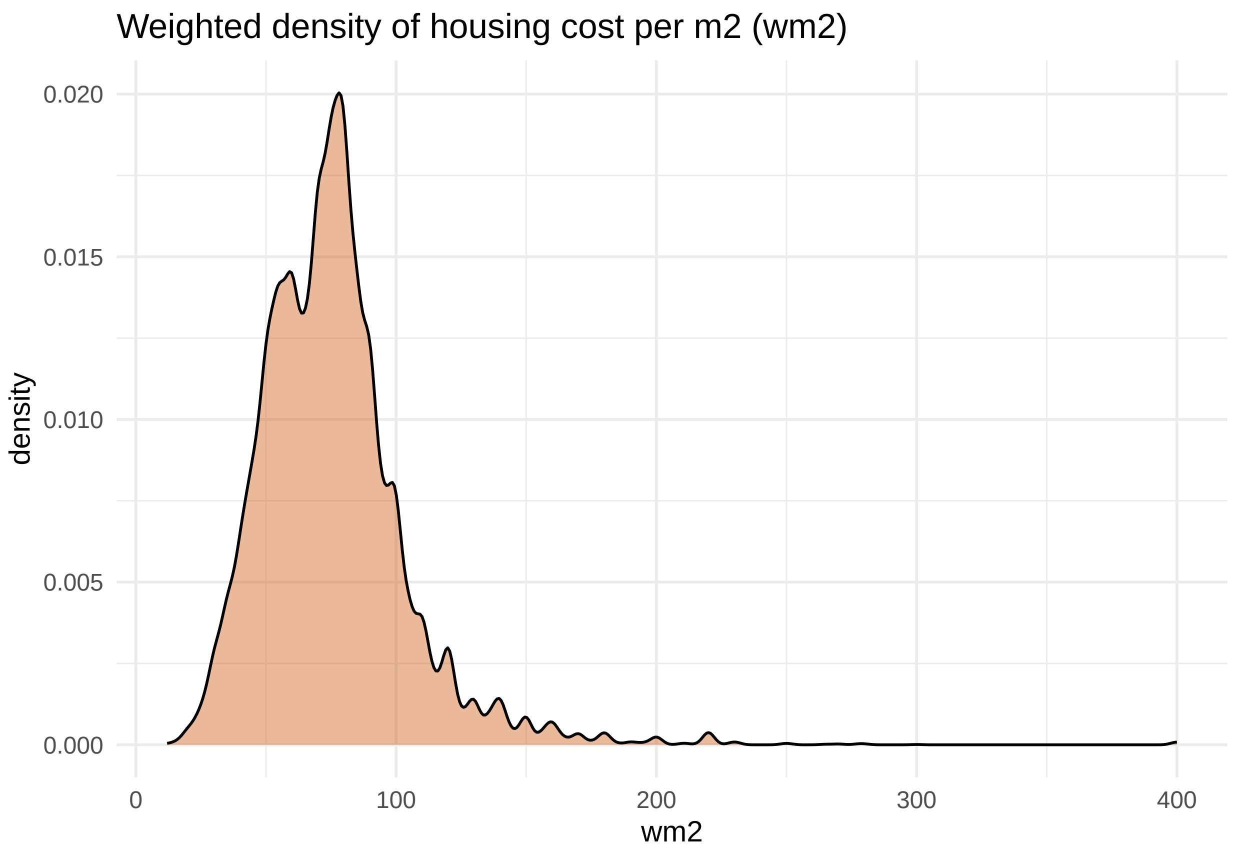
```
# Restrict sample to renting as main tenants using provided flags
# wrechtk == 1 (renter), wrechtl == 3 (main tenant)
renters <- df %>% filter(wrechtk == 1 & wrechtl == 3)

# Weighted density plot using ggplot with weights
```

```r
ggplot(renters, aes(x = wkges, weight = gewjahr)) +
  geom_density(fill = "#0c4c8a", alpha = 0.4) +
  labs(title = "Weighted density of total housing costs (wkges)", x = "wkges") +
  theme_minimal()
```



```r
ggplot(renters, aes(x = wm2, weight = gewjahr)) +
  geom_density(fill = "#c85102", alpha = 0.4) +
  labs(title = "Weighted density of housing cost per m2 (wm2)", x = "wm2") +
  theme_minimal()
```

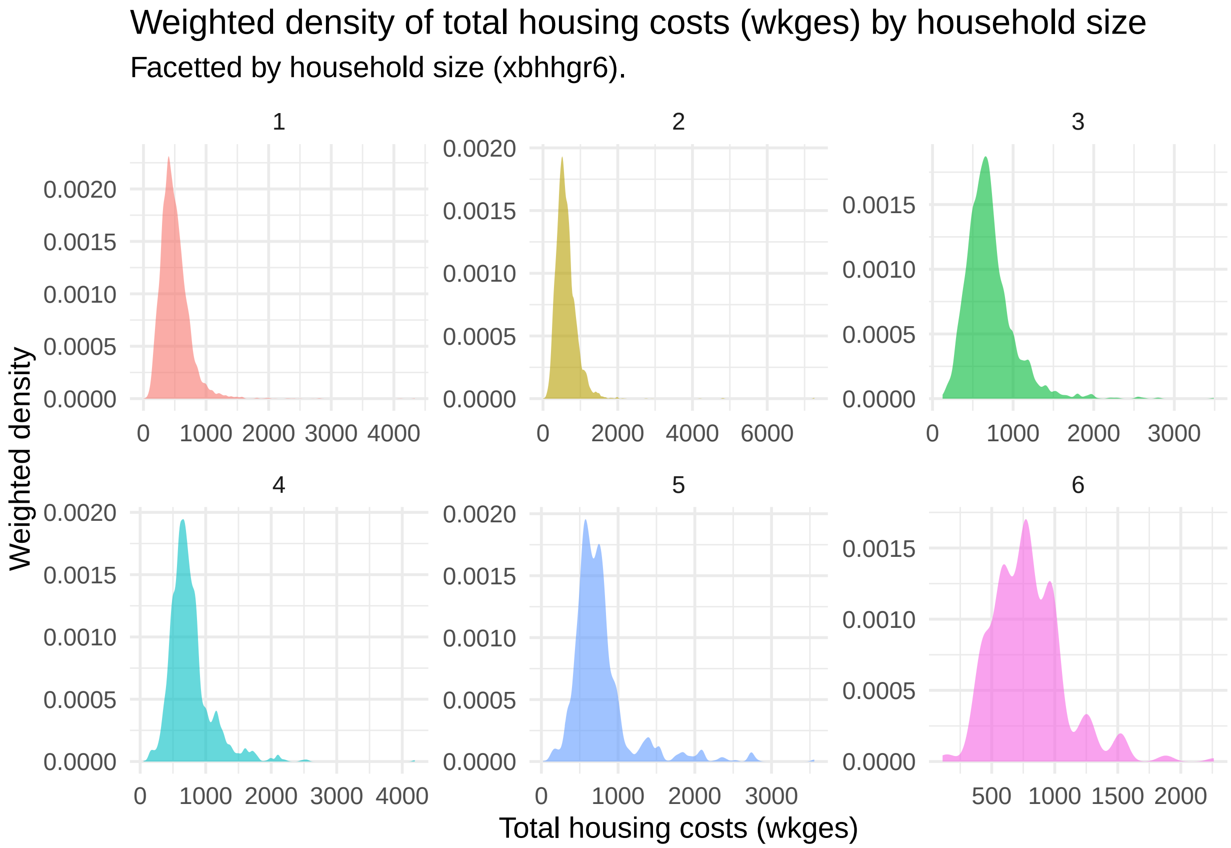


The first figure shows that most households have low total housing costs, with a few facing much higher expenses, indicating a right-skewed distribution and potential inequality. The second figure reveals that many households pay relatively little per square meter, while a minority face high spatial costs, suggesting urban concentration or market segmentation. Together, the plots highlight differences in absolute burden versus cost efficiency, helping distinguish between households that are economically stressed and those living in expensive but small spaces.

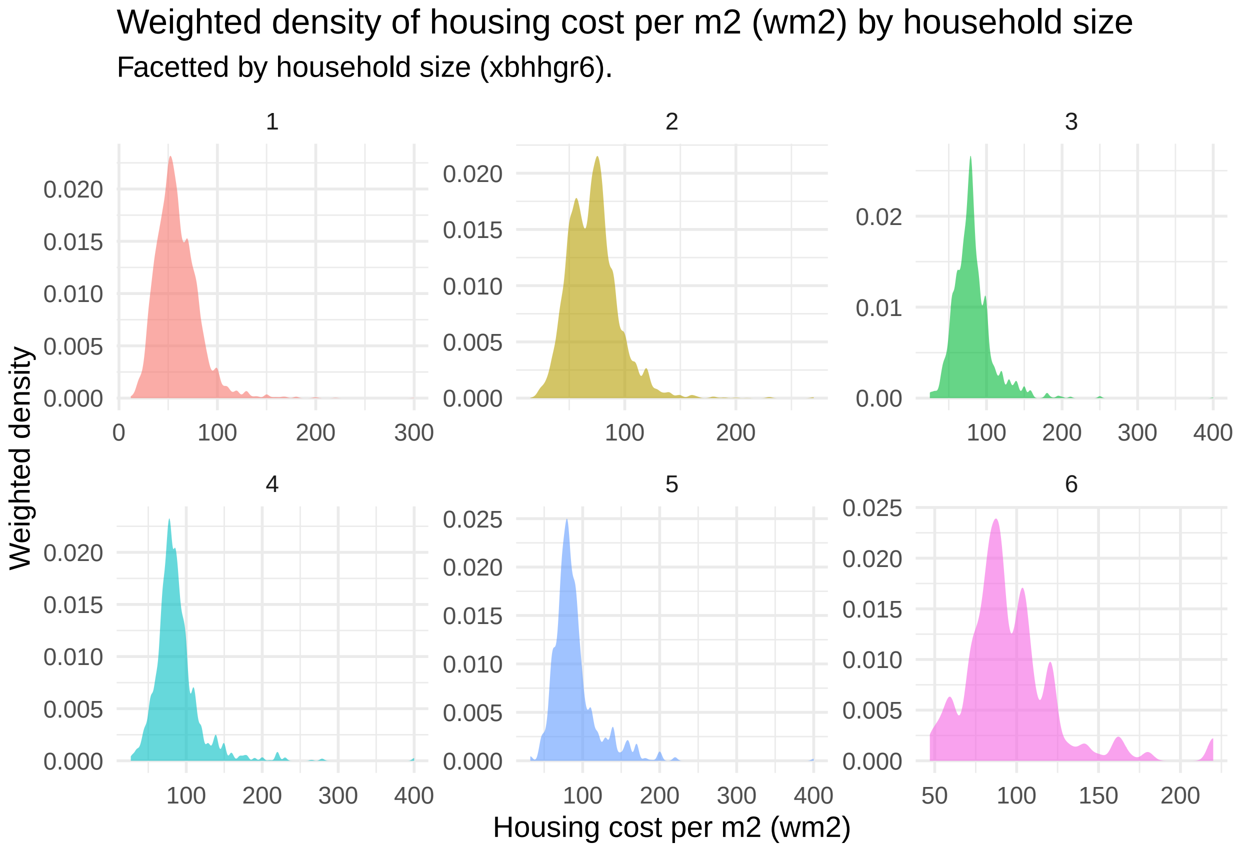
(c) Plot the weighted distributions of the two housing cost variables by two household characteristics of your choosing (present each separately); explain why you chose those characteristics and interpret the results.

```r
# Choose two household characteristics to group by
# xbhhgr6: household size in 6 categories (6 and more)
# xhhtyp2: household type
# (1 = single-family, 2 = two- and multi-family, 3 = non-family households)

# Weighted density of wkges by household size
renters %>%
  filter(!is.na(xbhhgr6) & !is.na(wkges) & !is.na(gewjahr)) %>%
  ggplot(aes(x = wkges, weight = gewjahr, fill = as.factor(xbhhgr6))) +
  geom_density(alpha = 0.6, color = NA) +
  facet_wrap(~ xbhhgr6, scales = "free") +
  labs(
    title = "Weighted density of total housing costs (wkges) by household size",
    subtitle = "Facetted by household size (xbhhgr6).",
    x = "Total housing costs (wkges)",
    y = "Weighted density"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```



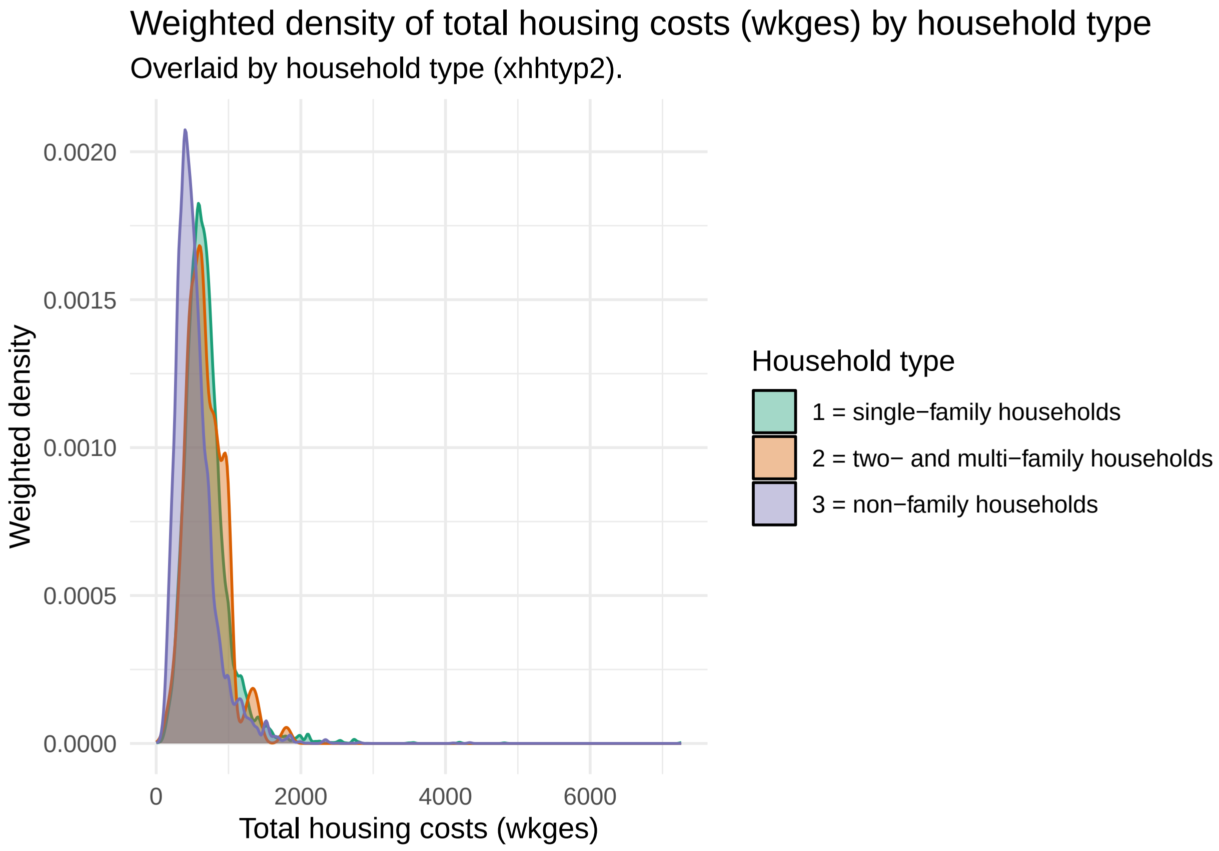
```
# Weighted density of wm2 by household size
renters %>%
  filter(!is.na(xbhhgr6) & !is.na(wm2) & !is.na(gewjahr)) %>%
  ggplot(aes(x = wm2, weight = gewjahr, fill = as.factor(xbhhgr6))) +
  geom_density(alpha = 0.6, color = NA) +
  facet_wrap(~ xbhhgr6, scales = "free") +
  labs(
    title = "Weighted density of housing cost per m2 (wm2) by household size",
    subtitle = "Facetted by household size (xbhhgr6).",
    x = "Housing cost per m2 (wm2)",
    y = "Weighted density"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```



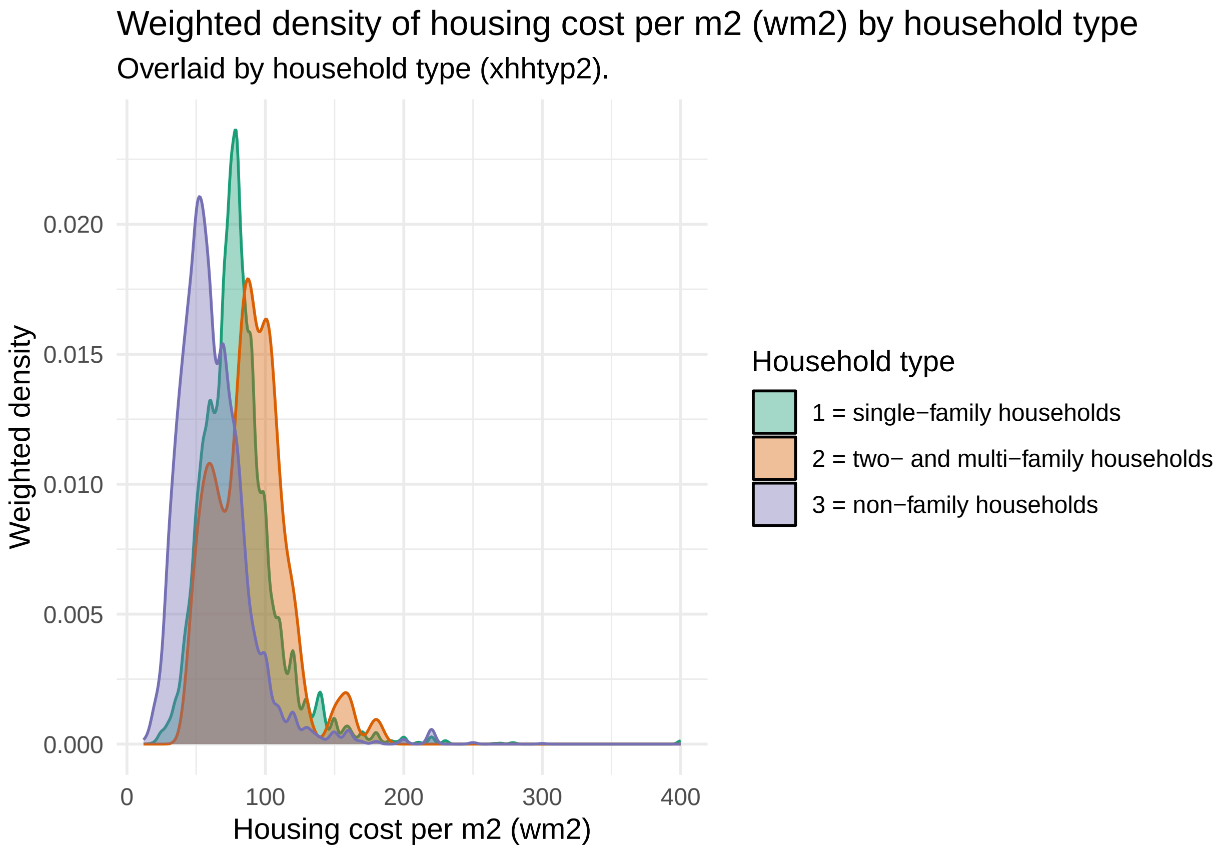
We choose *xbhhgr6* because household size directly influences housing needs and cost burdens. In both figures, the right-skewness of the distributions becomes more pronounced as household size increases. This means that while most larger households face moderate housing costs, a growing minority experience substantially higher costs — both in total (*wkges*) and per square meter (*wm2*). The increasing skew suggests rising inequality in housing affordability as household size grows.

We also choose *xhhtyp2* (household type) because it captures structural differences in living arrangements that influence housing costs — such as whether households are family-based, multi-family, or non-family.

```r
# Overlaid weighted density of wkges by household type (no facets)
renters %>%
  filter(!is.na(xhhtyp2) & !is.na(wkges) & !is.na(gewjahr)) %>%
  ggplot(aes(x = wkges, weight = gewjahr, fill = factor(xhhtyp2),
             color = factor(xhhtyp2))) +
  geom_density(alpha = 0.4, position = "identity") +
  scale_fill_manual(
    values = c("#1b9e77", "#d95f02", "#7570b3"),
    labels = c(
      "1 = single-family households",
      "2 = two- and multi-family households",
      "3 = non-family households"
    ),
    name = "Household type"
  ) +
  scale_color_manual(
    values = c("#1b9e77", "#d95f02", "#7570b3"),
    guide = FALSE
  ) +
  labs(
    title = "Weighted density of total housing costs (wkges) by household type",
    subtitle = "Overlaid by household type (xhhtyp2).",
    x = "Total housing costs (wkges)",
    y = "Weighted density"
  ) +
  theme_minimal() +
  theme(legend.position = "right")
```



```
# Overlaid weighted density of wm2 by household type (no facets)
renters %>%
  filter(!is.na(xhhtyp2) & !is.na(wm2) & !is.na(gewjahr)) %>%
  ggplot(aes(x = wm2, weight = gewjahr, fill = factor(xhhtyp2),
             color = factor(xhhtyp2))) +
  geom_density(alpha = 0.4, position = "identity") +
  scale_fill_manual(
    values = c("#1b9e77", "#d95f02", "#7570b3"),
    labels = c(
      "1 = single-family households",
      "2 = two- and multi-family households",
      "3 = non-family households"
    ),
    name = "Household type"
  ) +
  scale_color_manual(
    values = c("#1b9e77", "#d95f02", "#7570b3"),
    guide = FALSE
  ) +
  labs(
    title = "Weighted density of housing cost per m2 (wm2) by household type",
    subtitle = "Overlaid by household type (xhhtyp2).",
    x = "Housing cost per m2 (wm2)",
    y = "Weighted density"
  ) +
  theme_minimal() +
  theme(legend.position = "right")
```



The distributions of housing costs are right-skewed in both figures, but the second figure (housing cost per square meter, *wm2*) displays the differences more clearly. Single-family and non-family households tend to spend less at the mode, with single-family households showing the most concentrated distribution — indicating more uniformity in spatial costs. In contrast, multi-family households exhibit a more pronounced right skew, suggesting greater variability and a subset facing significantly higher per-unit costs. This pattern reflects how household structure shapes not just total housing burden but also the dispersion and inequality of spatial expenses.